단백질 상호작용 네트워크에서의 템플릿 기반 바이오 컴포넌트 탐색

박종민[O] 최재훈 박선희
한국전자통신연구원
{ jmpark93[O], jhchoi, shp}@etri.re.kr

Template-based Approach for Detecting Bio-Component in Protein Interaction Network

JongMin Park[O] JaeHun Choi SeonHee Park
Electronics & Telecommunications Research Institute(ETRI)

요 약

단백질 상호작용 네트워크에는 단백질들로 구성된 패스웨이와 콤플렉스 등과 같은 의미 있는 바이오 컴포넌트들이 존재한다. 하지만, 단백질 상호작용 네트워크는 방대한 단백질들과 상호작용 관계들로 구성되어 있고 많은 잘못된 정보들을 포함하고 있다. 따라서, 사용자가 정확한 단백질에 대한 식별자로 구성된 질의를 통해 원하는 바이오 컴포넌트를 탐색하는 것은 쉽지 않다. 본 논문에서는 사용자가 원하는 바이오 컴포넌트를 식별자뿐만 아니라 단백질 및 상호작용 관계의 다양한 특징들을 이용하여 탐색할 수 있는 방법을 제시한다. 또한, 단백질 상호작용 네트워크에는 잘못된 정보들이 많이 포함하고 있으므로 주어진 질의와 근접하게 일치하는 결과들도 제시할 수 있는 질의 연산자들을 제공하여 보다 다양한 관점에서 검토할 수 있도록 하였다.

1. 서 론

일반적으로 하나의 단백질은 고유한 기능을 가지고 있지만, 생체 내에서 특정한 생물학적 역할을 하기 위해서 여러 다른 단백질들과 다양한 상호작용을 한다. 따라서, 하나의 세포 내에는 많은 단백질들 사이에 복잡한 상호작용 관계들이 존재한다[1]. 이러한 단백질들 사이의 상호작용 관계들을 단백질은 노드로, 이들 사이의 상호작용을 링크로 표현하면 네트워크 형태로 나타낼 수 있다[2]. 또한, ' Hemoglobine' , ' DNA replication' 등과 같이 여러 개의 단백질들이 하나의 복합체(complex)를 구성하여 고유한 기능을 수행하기도 한다.

현재, 단백질 상호작용 네트워크는 보통 ' Yeast Two-Hybrid' 라는 생물학적 실험을 통해 빠르게 추출 되고 있으며, 추출된 상호작용 네트워크는 BIND(Biological Interaction Network Database), DIP(Database of Interacting Protein) 등과 같이 데이터베이스에 체계적으로 관리되고 있다. 또한, ' Mass Spectrometry' 실험 방법에 따르면 복합체를 식별할 수 있다고 한다. 하지만, 이러한 실험들을 통해 나오 결과들은 신뢰할 수 없는 정보들이 많이 포함되어 있다는 문제점이 있다.

본 논문에서 정의한 바이오 컴포넌트(Bio-Component)는 단백질 상호작용 네트워크에서 의미있는 단백질 패스웨이(pathway)나 복합체들을 말한다. 단백질 상호작용 네트워크 중에 사용자가 관심있는 질병 또는 대사 조절 과정 등에 참여하는 일부 바이오 컴포넌트들은 탐색하는 것은 실험 대상 단백질을 선정하거나 줄이는데 유용하게 사용될 수 있다. 또한, 특정 종에 나타난 바이오 컴포넌트들을 같은 종(species) 이나 다른 종에서도 탐색해서 비교해 봄으로써 바이오 컴포넌트의 정확성 및 완전성을 높일 수도 있다.

본 논문에서 제안한 템플릿 기반 바이오 컴포넌트 탐색 방법은 사용자가 원하는 바이오 컴포넌트를 식별자뿐만 아니라 단백질의 다양한 특징들을 이용하여 탐색할 수 있도록 하였다. 또한, 단백질 상호작용 네트워크에는 잘못된 정보들이 많이 포함하고 있으므로 주어진 질의와 근접하게 일치하는 결과들도 제시할 수 있는 질의 연산자들을 제공하여 보다 다양한 관점에서 검토할 수 있도록 하였다.

2. 관련연구

단백질 상호작용 네트워크에서 사용자가 원하는 바이오 컴포넌트를 식별하기 위해서는 전체 네트워크에서 주어진 그래프와 일치하는 부분 그래프를 찾을 수 있어야 한다. 부부 그래프 매칭(sub-graph matching) 방법은 크게 동형 매칭(isomorphism)과 준동형 매칭(homomorphism) 두 가지로 나눌 수 있다[3]. 먼저, 동형 매칭 방법은 그래프의 구조와 각 노드의 이름이 정확히 일치하는 그래프를 찾는 방법이다. 노드의 이름이 존재하는 경우 검색과정이 좀 더 쉬워지게 된다. 다음으로 준동형 매칭 방법은 그래프의 구조가 정확히 일치하는 그래프를 검색하지만 노드의 이름이 의미적으로 일치하는 경우에도 검색 대상으로 포함하여 처리한다.

본 논문에서 다루고 있는 단백질 및 상호작용 관계들은 이름뿐만 아니라 다른 특징들로도 표현될 수 있다. 따라서, 단백질과 상호작용 관계들을 매칭할 경우에는 의미적으로 일치하는 대상들도 포함하여야 한다. 또한, 단백질 상호작용

네트워크는 실험 방법 상의 문제로 많은 잘못된 정보들을 포함하고 있으므로, 사용자의 질의와 구조적으로 근접하게 일치하는 부분 그래프들도 제시할 수 있어야 한다. 또한, 단백질 상호작용 네트워크의 특성상 두개의 단백질 사이의 관계들로 구성되므로 일반적인 그래프 모델보다는 트리플 기반(triple-based) 모델을 활용한 검색방법이 적합하다.

3. 단백질 상호작용 네트워크와 바이오 컴포넌트

본 논문에서 다루는 단백질 상호작용 네트워크는 DIP, BIND 등에 공개된 이진 관계(binary relation) 데이터들로 구성되므로 다음과 같이 트리플 기반 모델로 네트워크를 표현하는 것이 자연스럽다. 여기서 정의한 트리플은 인스턴스 트리플(insance triple)이라 한다.

$$t = < r_{ij}, o_i, o_j > \; where \; o_i, o_j \in O \; and \; r_{ij} \in R$$

여기서, 단백질(o_i, o_j) 및 상호작용 관계(r_{ij})는 다음과 같은 속성을 가진다.

단백질 := Refer ID | Name | Gene | Annotation 상호작용 관계 := Name| Direction | Weight | Annotation

위에서 정의한 트리플이 여러 개 모여서 의미 있는 바이오 컴포넌트를 구성한다. 이러한 바이오 컴포넌트들은 실제 단백질 패스웨이와 콤플렉스 형태로 단백질 상호작용 네트워크에 나타난다. 아래 그림은 KEGG[4]에서 발췌한 파킨스 병(Parkinson's Disease)에 대한 일부 단백질 패스웨이 정보이다. ⇢ 표시는 상호작용 관계가 생략되었다는 것을 의미한다. 따라서, 기존의 동형, 준동형 매칭 방법으로는 검색할 수 없다. 또한, 각각의 노드에는 단백질 이름이 아니라 유전자(Gene)로 되어 있어서 실제 단백질 상호작용 네트워크에서 해당 바이오 컴포넌트를 검색하기 어렵다.

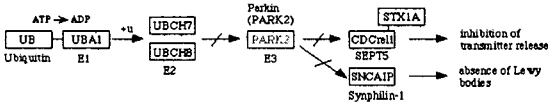


그림 1 파킨스 병에 대한 일부 단백질 패스웨이

아래 그림은 BIND[5]에서 발췌한 'p53' 단백질 비활성화에 관련된 콤플렉스이다. 3개의 구성 단백질들은 각각 고유한 기능을 가지고 있지만 하나로 모였을 때 다른 기능을 수행한다는 사실을 알 수 있다.

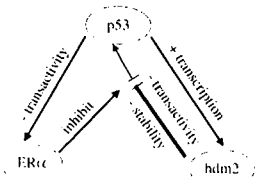


그림 2 단백질 콤플렉스 예

전체 단백질 상호작용 네트워크에서 단백질 패스웨이나 콤플렉스를 찾으면 이것들과 상호작용을 수행하는 주변 단백질을 알 수 있으며, 같은 종 또는 다른 종에서 이것과 유사한 근접 구조를 가지는 다른 부분 그래프들을 검색함으로써 비교 및 검증도 할 수 있으며, 전체 네트워크를 이해할 수 있도록 해준다.

4. 템플릿 기반 바이오 컴포넌트 탐색 방법

본 논문에서는 단백질 상호작용 네트워크를 인스턴스 트리플로 정의하였다. 정의된 트리플 집합 내에서 원하는 바이오 컴포넌트를 검색하기 위한 사용자 질의 또한 트리플로 구성한다. 이때 사용되는 트리플은 인스턴스 트리플을 확장한 템플릿 트리플(template triple)이다. 인스턴스 트리플에서는 실제 네트워크에 존재하는 단백질과 관계들로 정의되지만 템플릿 형태의 트리플에서는 단백질과 관계들에 대한 일반적인 용어들로 정의한다. 따라서, 하나의 템플릿 트리플은 실제 네트워크에서 여러 개의 인스턴스 트리플과 매칭된다.

$$ct = < cr_{ij}, c_i, c_j > \; where \; c_i, c_j \in CO \; and \; cr_{ij} \in CR$$
$$CO \; is \; set \; of \; term \; for \; protein,$$
$$CR \; is \; set \; of \; term \; for \; relation$$

예를 들어, 다음 그림에서 템플릿 트리플 $\alpha = < -, c_1, c_2 >$ 는 $\{< -, p_1, p_3 >, < -, p_2, p_4 >, < -, p_7, p_8 >\}$ 과 매칭되는 것을 알 수 있다. (여기서, c_1, c_2, c_2 는 개념 용어이다.)

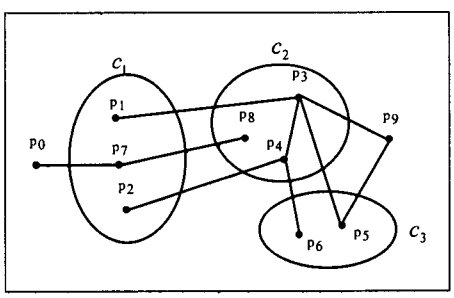
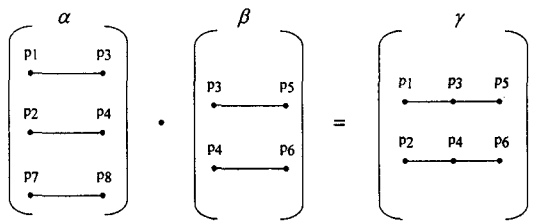


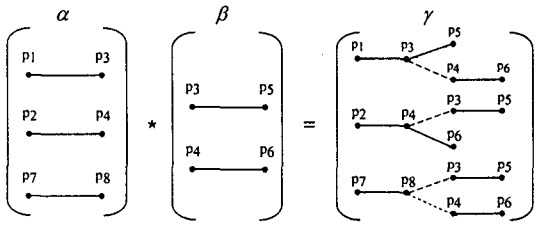
그림 3 단백질 상호작용 네트워크의 예

잘못된 정보를 대다수 포함하고 있는 단백질 상호작용 네트워크에서 사용자가 원하는 형태와 어느 정도 일치하는 바이오 컴포넌트를 검색하기 위해 템플릿 트리플 간에 사용되는 3가지 연산자를 정의하였다. 먼저, Association 연산자(•)는 질의에서 사용된 템플릿 트리플과 매칭되는 인스턴스 트리플 사이에 직접적으로 연관성이 존재하는 경우만을 검색할 때 사용한다. 다음으로, Cartesian 연산자(⋆)는 템플릿 트리플간의 직접적인 연결 관계에 상관없이 질의 트리플에 만족하는 모든 인스턴스 트리플을 검색한다. 따라서, 인스턴스 트리플 사이에 다른 상호작용 관계들이 포함된 바이오 컴포넌트도 검색할 수 있다. 마지막으로 Union 연산자(+)는 템플릿 트리플에 대한 인스턴스 트리플을 검색 결과에 모두 포함시킬 경우에 사용할 수 있다.

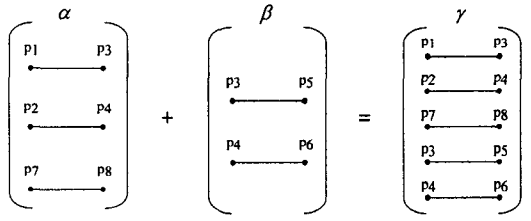
그림 3의 단백질 상호작용 네트워크에서 템플릿 트리플 α, β를 $\alpha = <-, c_1, c_2>$, $\beta = <-, c_2, c_3>$로 정의하였을 때 각 연산자를 적용한 예를 다음과 같이 나타내었다.



(a) $\gamma = \alpha \bullet \beta$



(b) $\gamma = \alpha * \beta$



(c) $\gamma = \alpha + \beta$

위에서 제시한 연산자를 적용하여 그림 1에 대한 질의를 템플릿 트리플로 구성하면 다음과 같이 표현할 수 있다. 여기서 개념 용어는 유전자(Gene)로 사용하였다.

<-, UB, UBA1>
• (<→, UBA1, UBCH7> + <→, UBA1, UBCH8>)
* (<→, BCH7, PARK2> + <→, UBCH8, PARK2>)
* (<→, PARK2, CDCrell> • <-, CDCrell, STX1A>)
* <→, PARK2, SNCAIP>

위와 같은 텍스트 질의를 구성하는 템플릿 트리플에서 사용되는 개념 용어를 실제 단백질들과 매칭할 때에는 단백질이 가지는 여러 특징들과 차례로 매칭을 수행하여 일치 정도 값을 계산한다. 특히, 개념 용어가 Annotation 인 경우 GO(Gene Ontology)[6]를 사용하여 실제 단백질이 가지는 Annotation 과의 개념 거리를 일치 정도 값으로 계산한다. 그리고, 사용자 인터페이스를 이용하여 텍스트 질의로 표현하기 힘든 복잡한 질의를 그래프 형태로 작성할 수 있다. 개념 용어를 질의 그래프에서 개념 노드로 작성하고 사용자가 원하는 세부적인 특징들을 지정할 수 있다. 예를 들어, 개념 노드에 생물학적 기능과 셀(Cell)에서의 위치를 동시에 지정하여 보다 세밀한 결과를 얻을 수 있다. 사용자 인터페이스를 이용한 질의는 실제 처리시에는 텍스트 형태의 질의 구조로 바꾸어 처리한다.

5. 템플릿 기반 바이오 컴포넌트 탐색 방법 설계 및 구현

본 논문에서 제안한 템플릿 기반 바이오 컴포넌트 탐색 방법은 기존의 단백질 상호작용 네트워크 관리 도구에 설계 및 구현되었다. 질의 처리시에 템플릿 트리플에 표현된 개념 노드는 실제 모든 상호작용 네트워크 나타나는 단백질들과 비교해야 되므로 많은 시간이 소요된다. 따라서, 개념 노드에 대한 인덱스를 생성하여 매칭과정에 소요되는 시간을 줄였다. 그림 4 는 대상 네트워크에서 바이오 컴포넌트를 탐색하고 결과 그래프와 연관성 있는 단백질을 포함하여 따로 분석하는 과정을 나타낸다.

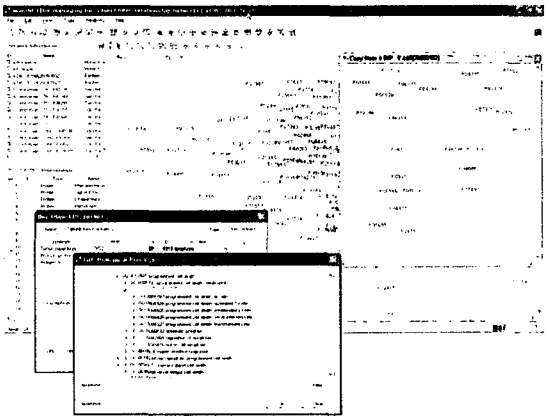


그림 4. 템플릿 기반 바이오 컴포넌트 탐색 화면

5. 결론 및 향후 연구 과제

본 논문에서 제안한 템플릿 기반 바이오 컴포넌트 탐색 방법은 사용자가 원하는 바이오 컴포넌트를 식별자뿐만 아니라 단백질의 다양한 특징들을 이용하여 탐색할 수 있도록 하였다. 또한, 단백질 상호작용 네트워크에는 잘못된 정보들이 많이 포함하고 있으므로 주어진 질의와 근접하게 일치하는 결과들도 제시할 수 있는 질의 연산자들을 제공하여 보다 다양한 관점에서 검토할 수 있도록 하였다.

현재 3가지 단순한 질의 연산자에 대해 정의하였지만 보다 사용자의 요구를 자연스럽게 표현할 수 있는 추가적인 질의 연산자에 대한 연구가 필요하며, 개념 노드와 실제 단백질간의 매칭시에 좀 더 정밀한 매칭 방법과 성능 개선을 위한 알고리즘에 대한 연구가 필요하다.

참고문헌

[1] C. L. Tucker, J. F. Gera, and P. Uetz, "Towards an Understanding of Complex Protein Interaction Maps.", Trends in Cell Biology, Vol. 11, No. 23, 2001.
[2] S. Oliver, "Guilt-by-Association Goes Global," Nature-News and Views, Vol. 403, 2000.
[3] B. Messmer and H. Bunke. "Efficient subgraph isomorphism detection - a decompostion approach.", To appear in IEEE Trans. on DKE, 2000.
[4] BIND, http://bind.ca/
[5] KEGG, http://www.genome.ad.jp/kegg/
[6] Gene Ontology, http://www.geneontology.org